

# Multiple Imputation in Survival Models: Applied on Breast Cancer Data

MR Baneshi<sup>1\*</sup>, AR Talei<sup>2</sup>

<sup>1</sup>Department of Biostatistics and Epidemiology, Kerman University of Medical Sciences, Kerman, Iran

<sup>2</sup>Department of Surgery, Shiraz University of Medical Sciences, Shiraz, Iran

## Abstract

**Background:** Missing data is a common problem in cancer research. While simple methods such as complete-case (C-C) analysis are commonly employed for handling this problem, several studies have shown that these methods led to biased estimates. We aim to address the methodological issues in development of a prognostic model with missing data.

**Methods:** Three hundred and ten breast cancer patients were enrolled. At first, patients with missing data on any of four candidate variables were omitted. Secondly, missing data were imputed 10 times. Cox regression model was fitted to the C-C and imputed data. Results were compared in terms of variables retained in the model, discrimination ability, and goodness of fit.

**Results:** Some variables lost their effect in complete-case analysis, due to loss in power, but reached significance level after imputation of missing data. Discrimination ability and goodness of fit of imputed data sets model was higher than that of complete-case model (C-index 76% versus 72%; Likelihood Ratio Test 51.19 versus 32.44).

**Conclusion:** Our findings showed inappropriateness of ad hoc complete-case analysis. This approach led to loss in power and imprecise estimates. Application of multiple imputation techniques to avoid such problems is recommended.

**Keywords:** Prognostic model; Missing data; Multiple imputation; Breast cancer

## Introduction

Prognostic models combine key patient characteristics (risk factors) to predict clinical outcomes such as recurrence of cancer. These models are excellent tools to investigate the contribution of variables to disease course, and to select the appropriate treatment paths.<sup>1</sup>

However, if in development of model, one ignores model assumptions, results might be misleading.<sup>2,3</sup> One of the issues that challenge the modelling practice is incomplete data. A problem in survival analysis occurs when data are missing on risk factors.<sup>4</sup> The traditional response to this problem is to exclude individuals with incomplete data on any prognostic factors from analysis (Known as Complete-Case Analysis (C-C analysis)).<sup>4</sup>

However, exclusion of missing data leads to attrition in sample size which will diminish precision of estimates and can lead to biased estimates.<sup>5,6</sup> Therefore, appropriate methods should be applied to impute missing data. Methodological developments in the field of analysis of missing data offers a lot to modelling. Advanced likelihood-based methods can be applied to use partially observed data so as to predict missing values. This preserves attrition in sample size and avoids biased estimates.

There are lots of methods to tackle the problem of missing data. The main aim of this paper is to highlight the methodological issues in development of a prognostic model in presence of missing data. Here we only focused on the Multivariable Imputation via Chained Equations (MICE) method. The MICE is a flexible method which has the capability to deal with all forms of variables (continuous, categorical, and

\*Correspondence: Mohammad Reza Baneshi, PhD, Department of Biostatistics and Epidemiology, Kerman University of Medical Sciences, Kerman, Iran. Tel: +98-913- 4423948, e-mail: m\_baneshi@kmu.ac.ir  
Received: January 10, 2011 Accepted: May 9, 2011

binary), and can be used in regression settings. Methods were applied analysing a breast cancer data set. To show the power of the MICE method in recovery of information, prognostic models were developed using complete data as well as imputed data sets.

## Materials and Methods

From 1994 to 2003, the information of 310 breast cancer patients in Shiraz (located in southern Iran) with a median follow-up of 2.5 years, were collected from Hospital-based Cancer Registry of Nemazee Hospital (affiliated to Shiraz University of Medical Sciences). The end point of the study was death. At the end of the study, there had been 56 deaths.

Variables offered to the multifactorial models were those showed to have univariate predictive ability (tumor stage with 3 levels (early, locally advanced, and advanced), tumor grade with 3 levels (1, 2 and 3), history of benign breast disease (positive versus negative), and age at diagnosis ( $\leq 47$  versus  $> 47$ ).<sup>7</sup> The data set do not include personal information such as name, address, or phone number of patients.

For analyzing of data, Kaplan-Meier and Log-rank tests were used to compare the survival curves in different groups. Linear Cox model was then applied to develop the multifactorial regression models and to estimate Hazard Ratios (HR).<sup>8</sup> Two models were used using Complete-Case (C-C) data and imputed data sets. To impute the missing data, Multivariable Imputation via Chained Equations were applied (MICE model).

In the C-C model, patients with missing data on any of 4 variables selected were excluded. Cox regression model in conjunction with ENTER variable selection method was then fitted to patients with available data on all 4 candidate risk factors. A final risk score was calculated by multiplying variables into the estimated regression coefficient. Tertiles of the risk score estimated were applied as cut off to categorise patients into low (L), intermediate (I), and high (H) risk groups.

The MICE method is a probabilistic approach. The usual practice to reflect the uncertainty about the true values of the missing data, is to replace each missing value by 10 values leading to 10 imputed data sets.<sup>9,10</sup> The process of the MICE method is described below:

To identify the mechanism of missing data, an indicator variable for each of variables which had missing data was created. For example, indicator variable for stage variable shows whether patient had missing

value or not. Patients with available data get a value of 1 while others get 0. The association between this indicator variable, showing stage missing and rest of variables were assessed applying Chi-Square test. When the missingness depends on observed variables mechanism, it is called Missing At Random (MAR).

It has been suggested that, for best imputation, the outcome variable should be included in the imputation model.<sup>11</sup> Therefore, patients' outcome and set of four risk factors were used in the MICE algorithm.

Polytomous and logistic regression were used to impute missing data for categorical (stage and grade) and binary data (age and benign disease history) respectively.

The MICE method involves no distributional assumption and can be used to impute missing data for continuous, categorical, and binary variables. To impute missing value on a variable which include missing data, say  $X_j$ , a regression model relates  $X_j$  to other variables in the imputation model. This regression model is then used to create imputed values by drawing from the posterior predictive distribution. Each predictor with missing values is considered in turn using the current imputed values for each of the other predictors.<sup>12</sup> The iteration process ends when all variables had been updated technical details are given in Appendix.<sup>13,14</sup> This entire process was repeated and the imputed values which are created at the 5<sup>th</sup> round were used as the first imputed data set. The whole processes were repeated 10 times to replace each missing data by 10 values, thus creating 10 data sets.<sup>12</sup> The standard algorithm imputes each incomplete column in the data from left to right. It is known that this issue (i.e. order of the variables) is essentially irrelevant to the results.

The creation of 10 data sets means there is a requirement for 10 modelling analyses, one for each data set, and there will therefore be 10 different estimates for each parameter. Estimates derived from imputed data sets (the coefficients and standard errors) therefore, need to be combined and this was achieved applying Rubin's rule.<sup>14</sup> The final regression coefficient is simply the average of coefficients across imputed data sets.<sup>14</sup> In estimation of standard errors, both between and within imputation variations should be taken into account technical details are given in Appendix.

Hazard Ratios (HR) and corresponding 95% Confidence Intervals (C.I.) were calculated from regres-

sion coefficients and standard errors that have been imputed across multiply imputed data sets.

A risk score was calculated for each of 10 imputed data sets. For each patient, a single averaged risk score was calculated by averaging her estimated risk scores from each of the 10 imputed data sets.

In risk stratification studies, it is important to create risk groups where patients in each group are equally likely to develop the outcome.<sup>15</sup> Discrimination refers to the ability to separate patients with different responses<sup>15</sup> and is measured using Harrell's C-index (concordance index) which is a generalisation of Area Under Curve (AUC).<sup>16,17</sup> The C-index is interpreted as correct ordering in the sense that comparing risk predictions for two patients, risk calculated for whom developed the disease is higher than the other one. This statistic varies between 0.5 and 1 where values near 1 indicate high discrimination power. However, if performance is assessed on the same sample as used for model development, then performance will be overestimated. Therefore, bootstrap procedure was applied and bias-corrected C-indices were reported.<sup>18</sup>

For all models, we will report Likelihood Ratio Test (LRT) which indicates how well the model fits the data. A series of packages which work under R software (version 2.5.1) were used.<sup>19</sup> Missing data were imputed using MICE package. Estimated regression coefficients and standard errors were combined across imputed data sets using Mitools library. Performance of models (discrimination and predictive ability) was assessed using Design library. K-M curves are plotted using SPSS software.

## Results

Information for age variable was available for all patients. The variables nodal status and grade involved about 20% missing rate (20.3% and 20.6% respectively). Corresponding figure for 'history of benign disease' was 15.2%. However, after exclusion of missing data on all four variables, 35% of data were lost. Totally, 203 cases (65%) had data available on

all 4 variables. Almost all patients with missing data were those survived. Out of 56 deaths only 2 ones were lost in complete-case analysis.

We first examined missing data mechanism (Table 1). As shown, patient's status, grade and history of benign disease can predict missingness on stage variable. Patients' status and history of benign disease were predictors of grade missing. Furthermore, patients' status and age at diagnosis were predictors of benign disease variable. This confirms that data had a Missing At Random (MAR) mechanism.

Estimated Hazard ratios (HR) with 95% Confidence Intervals (C.I.), corresponding to complete-case and imputed data sets are given in Table 2. Age at diagnosis and history of benign disease were not significant in complete-case model, due to attrition in sample size and inevitable loss in power of model. Furthermore, risk of death for patients with stage 3 relative to those with stage 1 was not significant in complete-case model. After imputing missing data, both of these variables (age at diagnosis and family history of benign disease) were retained in the model. In addition, HR for cases with stage 3 relative to stage 1 reached a significance level.

Comparing performance of models, imputation of missing data led to 4 percentages point improvement in discrimination ability of model (76% for the MICE versus 72% for C-C data). Furthermore improvement in model goodness of fit was seen (51.19 versus 32.44).

## Discussion

Missing data are a common problem in medical and epidemiological data sets. Exclusion of missing data leads to loss of power. In results presented, some variables lost their significant effect in complete-case analysis. For example, stage of disease is known as one of the most important prognostic variables.<sup>20,21</sup> However, this variable did not reach to the significant level in the C-C model.

On the other had, in order to protect against chance effects dues to imputation, we imputed 10 data sets. This protection was to be felt worth the

**Table 1:** Investigation of the association between variables' missingness and the rest of variables<sup>a</sup>

Missing indicator	Status	Stage	Grade	Benign disease	Age
Stage	+		+	+	-
Grade	+	-		+	-
Benign disease	+	-	-		+

<sup>a</sup> +: association between missing indicator and variable, -: lack of association between missing indicator and variable

**Table 2:** Comparison of estimated HRs (95% C.I.s) corresponding to analysis of complete-case and imputed data sets<sup>a</sup>

Variable	Level	Complete-case model (N=203, D=54)		Imputed data sets model (N=310, D=56)	
		HR (95% C.I.)	P value	HR (95% C.I.)	P value
Stage	1	1		1	
	2	2.89 (1.52, 5.51)	0.001	3.13 (1.64, 5.97)	<0.001
	3	1.94 (0.81, 4.63)	0.13	2.53 (1.05, 6.12)	0.03
Grade	1	1		1	
	2	2.46 (1.61, 5.23)	0.02	2.46 (1.15, 5.24)	0.02
	3	1.33 (0.58, 3.04)	0.50	1.52 (0.65, 3.60)	0.34
Age	<48	1		1	
	≥48	1.75 (0.91, 3.38)	0.10	1.92 (1.01, 3.65)	0.04
Benign	No	1		1	
	Yes	1.91 (1.04, 3.49)	0.04	2.32 (1.24, 4.33)	0.01
Performance of models					
C-index		72%		76%	
Likelihood Ratio Test		32.44		51.19	

<sup>a</sup> HR: Hazard Ratio, C.I.: Confidence Interval, N: Sample size, D: Number of deaths

inconvenience of having to average risk scores across 10 final models. Once missing data were imputed, power was increased and variables lost their effect in complete-case model (such as stage of disease) and reached a significance level.

We also showed that our data had a Missing At Random (MAR) mechanism. This means that missing data depends on other patients' characteristics and therefore can be well imputed using multiple imputation methods. We should emphasize that our main goal was to illustrate the process of development of a prognostic model when missing data exist. To achieve this, we simply used a breast cancer data set in southern Iran as a prevalent cancer in this region set as an example.<sup>22,23</sup> Discussion of risk factors of breast cancer is beyond the scope of this paper and were previously reported.<sup>22,23</sup> This issue has been addressed here.<sup>20,21</sup>

It should be noted that, when missing rate is low, results of C-C model, in terms of variables retaining in the final model, might be similar to that of MICE. Asia Pacific Cohort Studies Collaborators (APCSC) collects data to determine Coronary Heart Disease (CHD) risk factor. Ability of multiple imputation and complete-case analysis to handle the missing data on a single variable (cholesterol) in 26 studies was compared.<sup>24</sup> Cholesterol missing rate varied from 0% to 69%. In 22 studies where cholesterol value was not available for about 10% of subjects both methods gave similar results. On the other hand, in four studies with missing rate between 10% to 60%, clear differ-

ence was seen between models. It has also been commented that with more than 60% missing rate, the MICE model might not provide accurate estimates.<sup>24</sup>

However, we believe that a low rate of missing data on each variable might cause serious problems in multivariate modelling when patients with missing data on different variables are not the same because this might substantially reduce the number of complete cases available for analysis, and increase the chance of bias due to excluded cases.

We developed the multifactorial models in conjunction with ENTER variable selection method. When Backward Elimination (B.E.) variable selection is hired, a series of iterative steps are required to exclude variables which do not contribute significantly to the model. If a single multifactorial model was developed, then application of B.E. is straightforward. However, when there are 10 imputed data sets, B.E. will not directly be feasible. In an iterative process, at each step, the results were aggregated across the 10 data sets, and the variable with the highest *P*-value (exceeding 0.05) was removed. Another set of 10 models were fitted with remaining variables, results were aggregated, and *P*-value assessed for a variable to drop (if *p*-value > 0.05). The whole process continued until all variables remained significant.<sup>12,13</sup>

Before development of multifactorial models, we dichotomised the variable age at 48 because we showed that dichotomised version of this variable, in comparison with continuous form, improved the quality of the model.<sup>25</sup> Therefore in this study, only

information on 2 binary and 2 categorical variables were analysed. When continuous data are available, Predictive Mean Matching (PMM) technique can be employed. In the PMM method, the complete-case whose value is closest to the imputed value is chosen. It takes the observation from the complete-case as the imputed value.

Our work involved several limitations. We used a data set contained only four variables. Therefore, impact of number of variables offered to the multifactorial model was not investigated. Furthermore, we only compared performance of the C-C and the MICE at 35% missing rate and under MAR mechanism. It is known that performance of models depends to a great extent to mechanism of missing data, rate of missing data, method of imputation of missing data, and sample size.<sup>26-29</sup> Our work was simply a case study to explain the methodological issues in the application of the MICE method, and its art in recovery of information.

Therefore, it is needed to design future studies so as to compare the performance of imputation models under different scenarios (i.e. by changing the sample size, missingness mechanism, missing rate, and method of imputation). We already showed that the C-C model decreases the power and the MICE method recovers the data. However, at this stage, due to limitations listed above, we cannot provide a

specific guideline on how best to tackle the problem of missing data because there are lots of approaches to deal with missing data.<sup>30</sup> It has been shown that under special circumstances, alternative methods with easier methodology (such as replacement of missing data by mean of observed values) might provide comparable estimates. Application and comparison of alternative imputation methods were beyond the scope of this paper and will be published elsewhere.

Results presented showed how exclusion of missing data affect the composition of the model. Application of ad hoc methods such as complete case analysis is hugely criticised.<sup>31,32</sup> When complete-case gives results comparable to that of the MICE method, a gold standard such as the MICE method is required to compare results with other simpler methods. Therefore, application of such methods is highly recommended.

## Acknowledgment

We should thank staff of Motahhari Paraclinic and Shahid Faghihi Hospital who facilitated our access to patients' folder and information.

**Conflict of interest:** None declared.

## References

- Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Res Treat* 1998; **52**:289-303. [10066088] [http://dx.doi.org/10.1023/A:1006193704132]
- Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med* 1993; **118**:201-10. [8417638]
- Wyatt JC, Altman DG. Prognostic models: clinically useful or simply forgotten. *Br Med J* 1961 Dec 9; **2**:1539-41. [13868802]
- Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br J Cancer* 2004; **91**:4-8. [15188004] [http://dx.doi.org/10.1038/sj.bjc.6601907]
- Altman DG, Bland JM. Missing data. *BMJ* 2007; **334**:424. [17322261] [http://dx.doi.org/10.1136/bmj.38977.682025.2C]
- Baneshi MR, Talei AR. Impact of imputation of missing data on estimation of survival rates: an example in breast cancer. *Iranian Journal of Cancer Prevention* 2010; **3**:127-31.
- Rajaeeafard AR, Baneshi MR, Talei AR, Mehrabani D. Survival Models in Breast Cancer. *Iran Red Crescent Med J* 2009; **11**:295-300.
- Cox DR. Regression models and life tables. *J R Stat Soc* 1972; **34**:187-220.
- Schafer JL. Analysis of Incomplete Multivariate Data. Florida: *Chapman and Hall*; 1997.
- Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res* 1999 Mar; **8**:3-15. [10347857] [http://dx.doi.org/10.1191/096228099671525676]
- Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006; **59**:1092-101. [16980150] [http://dx.doi.org/10.1016/j.jclinepi.2006.01.009]
- Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999; **18**:681-94. [10204197] [http://dx.doi.org/10.1002/(SICI)1097-0258(19990301)18:6<681::AID-SIM71>3.0.CO;2-R]
- Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J Clin Epidemiol* 2003; **56**:28-37. [12589867] [http://dx.doi.org/10.1016/S0895-4356(02)00539-5]
- Rubin DB. Multiple imputation for non response in surveys. 1978.
- Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999; **130**:515-24. [10075620]
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29-36. [7063747]
- Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004; **23**:2109-23. [15211606] [http://dx.doi.org/10.1002/sim.1802]
- Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for lo-

- gistic regression analysis. *J Clin Epidemiol* 2001;**54**:774-81. [11470385] [http://dx.doi.org/10.1016/S0895-4356(01)00341-9]
- 19 R: A language and environment for [statistical computing computer program]. 2007.
  - 20 Baneshi MR, Warner P, Anderson N, Tovey S, Edwards J, Bartlett JM. Can biomarkers improve ability of NPI in risk prediction? a decision tree model analysis. *Iran J Can Prev* 2010;**2**:62-74.
  - 21 Baneshi MR, Warner P, Anderson N, Bartlett JSM. Tamoxifen resistance in early breast cancer: statistical modelling of tissue markers to improve risk prediction. *British J Can* 2010;**102**:1503-10. [20461093] [http://dx.doi.org/10.1038/sj.bjc.6605627]
  - 22 Rajaeefard AR, Baneshi Mr, Talei Ar, Mehrabani D. Survival models in breast cancer patients. *Iran Red Crescent Med J* 2009;**11**:295-300.
  - 23 Mehrabani D, SZ Tabei, ST Heydari, SJ Shamsina, N Shokrpour, M Amini, SJ Masoumi, H Julaei, M Farahmand, A Manafi. Cancer occurrence in Fars Province, Southern Iran. *Iran Red Crescent Med J* 2008;**10**:314-322.
  - 24 Barzi F, Woodward M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *Am J Epidemiol* 2004;**160**:34-45. [1522911] [http://dx.doi.org/10.1093/aje/kwh175]
  - 25 Baneshi MR, Talei AR. Dichotomisation of continuous data: review of methods, advantages, and disadvantages. *Iranian Journal of Cancer Prevention* 2010;**4**:26-32.
  - 26 Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Stat Methods Med Res* 2007;**16**:277-98. [17621472] [http://dx.doi.org/10.1177/0962280206074466]
  - 27 Bono C, Ried LD, Kimberlin C, Vogel B. Missing data on the Center for Epidemiologic Studies Depression Scale: a comparison of 4 imputation techniques. *Res Social Adm Pharm* 2007;**3**:1-27. [17350555] [http://dx.doi.org/10.1016/j.sapharm.2006.04.001]
  - 28 Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;**59**:1087-91. [1698014] [http://dx.doi.org/10.1016/j.jclinepi.2006.01.014]
  - 29 Musil CM, Warner CB, Yobas PK, Jones SL. A comparison of imputation techniques for handling missing data. *West J Nurs Res* 2002;**24**:815-29. [12428897] [http://dx.doi.org/10.1177/019394502762477004]
  - 30 Baneshi MR. Statistical Models in Prognostic Modelling of Many Skewed Variables and Missing Data: A Case Study in Breast Cancer (PhD thesis submitted at Edinburgh University) 2009.
  - 31 Croy CD, Novins DK. Methods for addressing missing data in psychiatric and developmental research. *J Am Acad Child Adolesc Psychiatry* 2005;**44**:1230-40. [16292114] [http://dx.doi.org/10.1097/01.chi.0000181044.06337.6f]
  - 32 Van Der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J Clin Epidemiol* 2006;**59**:1102-9. [16980151] [http://dx.doi.org/10.1016/j.jclinepi.2006.01.015]
  - 33 Van Buuren S, Oudshoorn K. Multiple imputation by chained equations: MICE V1.0 User's manual. 2000.